Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 1 (2012) 372 – 377

Procedia Technology

INSODE 2011

Farsi Word Spotting and Font Size Recognition

Yaghoub Pourasad [a] *, Houshang Hassibi [b], Azam Ghorbani [c]

[a] Department of electrical and computer engineering, K. N. Toosi University of Technology, Tehran, Iran
[b] Department of electrical and computer engineering, K. N. Toosi University of Technology, Tehran, Iran
[c] Department of engineering, Saveh branch, Islamic Azad University, Saveh, Iran

Abstract

This paper is the first paper about Farsi word spotting and font size recognition. In this work using some font size independent features such as the number, aspect ratio, and mesh features of sub words, a Farsi keyword is described, searched and found. Also font size of document image is recognized. This approach has been evaluated on a dataset consisting of 500 Farsi document images where font size recognition rate of 94.2%. Retrieval precision rate of 92.3% at recall rate of 76.5% has been obtained. This approach with little adaptation is applicable on Arabic and Urdu documents.

Keywords: Farsi document image, Second word spotting, Font size recognition, Mesh features;

1. Introduction

There are many text documents which have been scanned and stored in digital libraries without OCR (Optical Character Recognition). There are two scenarios for information retrieval from these imaged documents. In a traditional retrieval scenario, document images are converted to the text through OCR [1]; but OCR softwares can't always transcribe document images to ASCII texts precisely. In fact when document images are in degraded quality and some adjacent characters are touching each others, performance of OCR softwares falls. Also for a huge amount of document images archived in digital libraries, OCR technique requires very long time. To overcome these problems researchers have proposed another way which is called keyword spotting. In word spotting methods searching is done directly in image domain instead of text domain. Most of the word spotting papers are presented for English (Latin) [2, 3] language and some of them are for other languages such as Chinese [4], Korean [5], Arabic [6], etc. There are few papers about Arabic word spotting; and most of these papers are about handwritten Arabic documents. We haven't seen any paper about Farsi document images. For example in [7] an algorithm and a system for searching handwritten Arabic documents is presented. In [8] a system for spotting words in scanned document images in three scripts, Devanagari, Arabic, and Latin is described. In [9] a system for searching keywords in Arabic handwritten historical documents using two algorithms, Dynamic Time Warping (DTW), and Hidden Markov Model (HMM), has been presented.

Font recognition is an important process which can be very useful in every OCR, spotting and retrieval system. Although there has been great attempts in producing Omni-font OCR systems for Farsi/Arabic language [10], the

* Yaghoub Pourasad. Tel.: +98 21 8888 2991
E-mail address: y_pourasad@ee.kntu.ac.ir

doi:10.1016/j.protcy.2012.02.077

overall performance of such systems are far from perfect. In spite of this importance, only few researchers have addressed font recognition issue. In recent years some papers about Farsi [11, 12, 13] and Arabic [14] font recognition have been reported which all of them are font size independent. These works recognize only font face of document images but don't recognize their font sizes. This paper is the first paper about Farsi word spotting and font size recognition. In this approach when a user enters a Farsi keyword through GUI, for searching in a document image, first, document's font face is recognized, then entered query word is rewritten and modified according to document's font face; and its image is constructed. In the next step, bounding boxes of modified word image is constructed and some font size independent features such as aspect ratio, mesh features, and number of its sub words are extracted for its description. These features are searched in the document image and the same word instances are found. After finding same query word instances in document, with comparing width and height of entered keyword with width and height of found words from document, font size of document is determined.
This paper is organized as follows. In section 2 our proposed method is described, in section 3 experimental results is presented and finally section 4 is conclusion.

2. Proposed Method

While presenting a word spotting method for a special language, its special characteristics should be considered. Different scripts have different characteristics and therefore different word spotting methods. Farsi scripts have some characteristics which make them different from English scripts; Farsi scripts are cursive in both handwritten and typewritten documents. In English documents, words are composed of disjoint characters with almost same sizes; whereas in Farsi documents, words are composed of some sub words instead of characters. Each sub word is composed of one or more connected characters. Although these characteristics lead to difficulties in text line segmentation to words or characters, but they can be good features for retrieval purposes. In fact number of sub words and their widths and heights in each word differ with other words. In figure1 some Farsi words and their sub words have been shown.

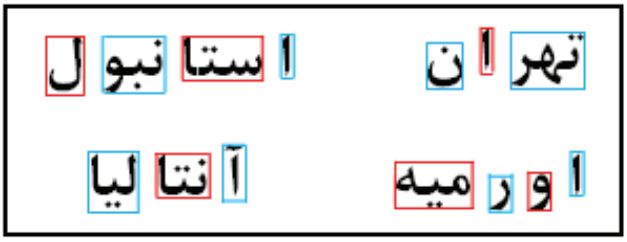

Fig. 1. some Farsi words and their sub words

There is another major difference between Farsi and English scripts. In English scripts, only two letters ('i' and' j') are composed of two components, whereas in Farsi, a large number (more than 19 out of 32) of letters are composed of more than one component. In figure 2 some Farsi letters which have two or more components have been shown.

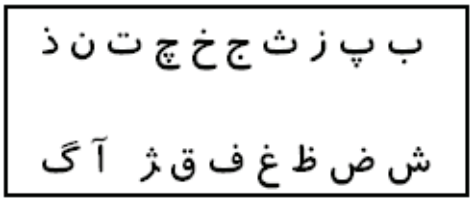

Fig. 2. Some Farsi letters which have two or more components

There are many structural and statistical features which have been used in word spotting literature; but we use some of them which are font size invariant. The first feature which we use in our approach is the number of sub

words of query word and their widths and heights. It should be noted that we use aspect ratio of sub words of words instead of their widths and heights so that used features be font size independent. The second used feature is the number of components in each sub word; and the third feature is mesh feature.

In the works like [5] which use mesh features, the image is divided logically into a fixed number of zones and for each zone the black pixel density is calculated, i.e., total number of ink pixels. The length of the feature vector is equal to the total number of zones. This type for division for Farsi words isn't always relevant; so we didn't extract mesh features for words in a fixed number of zones. We extracted mesh features for each sub word with adaptive divisions. In our method the rule for division is that a sub word with height of 'h' and width of 'w' will be divided only in its greatest direction. For example if h < w, then bounding box of sub word will be divided only horizontally from left to right; and width of each zone will be 'h'. In some cases width of the last zone may be smaller than 'h'. If w < h, then bounding box of each zone will be divided only vertically from top to down; and height of each zone will be 'w'. In some cases the height of the last zone may be smaller than 'w'. In figure 3 this process has been shown. This way of mesh feature extraction is robust and font size independent.

In order to match and search extracted features of query word through document image to find similar cases, we use a multi level matching and pruning process. In the first step, number of sub words of query word and their aspect ratios are considered and the same instances in document image are found. The next steps of matching are applied only on found word instances. This work reduces computational operations; because with applying each type of features, many irrelevant words are put away; and searching domain becomes smaller and smaller. It should be noted that in order to avoid of missing correct instances which are slightly different with query word, in each step of matching and pruning, a value as tolerance is considered. This work increases the recall rate of spotting process. After applying number of sub words and their aspect ratios features, number of components in each sub word and finally mesh features are applied. All words from document image which all their mentioned features correspond with query word features, are found and considered as correct answers.

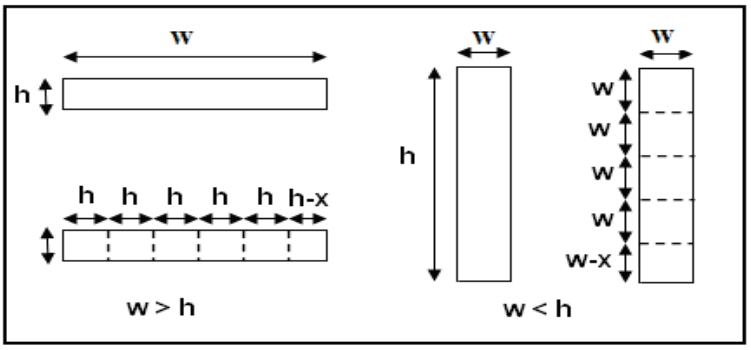

Fig. 3. Bounding box division for mesh feature extraction

After finding same instances of query word, we can recognize the font size of text of document image, with comparing width and height of query word, with width and height of same word instances in document image. The fact is that if the height of a sub word which is written with font size f1,is 'h1' and its width is 'w1' and its height when is written with font size f2 is 'h2' and its width is 'w2', equations (1) and (2) are true:

$$w1/w2 = f1/f2 \quad (1)$$

$$h1/h2 = f1/f2 \quad (2)$$

In practice, there are some problems such as 'Justifying' (Aligning text to both the left and right margins) which may disturb equation (1); but equation (2) is always correct; because the problems which change sub words width, don't change their height. In figure 4 two normal words and their justified form are shown.

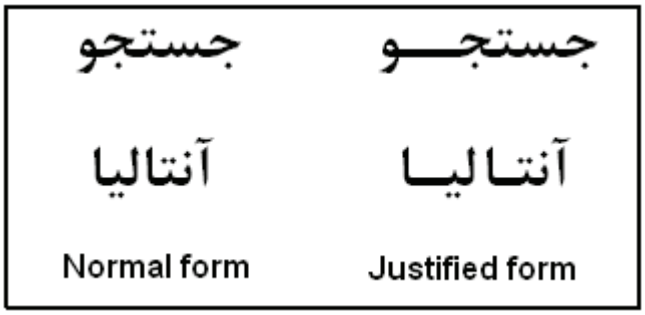

Fig. 4. Two Farsi words in normal and justified forms

For a query word and its same instance in the document image, in order to font size recognition, we can use equation (2) considering h1 as height of query word and h2 as height of same instance word in the document image; but experiments show that in order to increase the precision of font size recognition it is better that we consider height of all sub words of query word instead of only own word's height:

$$f2 = \frac{1}{n}\sum_{i=1}^{n}\frac{h1_i}{h1_i} * f1 \qquad i=1,2...n \tag{3}$$

$$f2 = f1 * \frac{1}{n}\sum_{i=1}^{n}\frac{h1_i}{h1_i} \qquad i=1,2,...n \tag{4}$$

Where n is the number of sub words.
In this work when query word is modified according the font face of document image, it is written with font size of 14; so:

$$f2 = 14 * \frac{1}{n}\sum_{i=1}^{n}\frac{h1_i}{h1_i} \qquad i=1,2,...n \tag{5}$$

3. Experimental Results

In order to evaluate our approach, we constructed a dataset consist of 500 Farsi document images. These documents are noiseless and without skew and are written in 10 common Farsi font faces (Lotus, Nazanin, Mitra, Yaghut, Koodak, Homa, Tahoma, Titr, Nasim, and Zar), in font sizes 8 up to 20. Also we provided a GUI (Graphical User Interface) so that we can enter
a Farsi word by it. Our method has been implemented and evaluated using MATLAB software, with a dual core, 2.4 GHz Pentium with 512 MB RAM memory. Common criteria for performance evaluating of spotting methods are Precision (P), Recall (R), and F1, which are defined as:

$$P = \frac{No\ of\ correctly\ \ detected\ \ keywords}{No\ of\ all\ detected\ \ words} \tag{6}$$

$$R = \frac{No\ of\ correctly\ \ detected\ \ keywords}{No\ of\ actual\ \ keyword\ \ appearanc\ \ es} \tag{7}$$

$$F1 = \frac{2*P*R}{P+R} \tag{8}$$

After entrance of a query word through GUI, in order to modification of it, according to font face of document image, font face of document image should be recognized. This work has been done using Khosravi and Kabir method [11] which is one of the fastest and the most precise font recognition methods. In order to extract sub words of query word and document image words, we used vertical projection profile.

Among all introduced features we first applied number of sub words and their aspect ratios. In order to compensate slight variations and also 'justifying' (Aligning text to both right and left margins) problem, we

considered a tolerance value (T) for aspect ratio of sub words in each word. Whatever tolerance value is increased, recall rate increases but precision rate decreases and vice versa. In table1 performance of system for 10 values of tolerances (T) are presented. In this case the best performance (best value of F1) is achieved with tolerance value of 12%. These results have been obtained with testing the approach on 200 words as queries. In this work considered value as tolerance of mesh features was 15%.

Table 1. Performance of spotting method for different values of ±T

T	2	4	6	8	10	12	14	16	18	20
P (%)	98.4	98.1	96.3	95.2	93.8	92.3	89.4	87.6	84.3	80.4
R (%)	67.7	69.1	71.4	72.1	73.1	76.5	77.2	78.4	80.5	82.6
F1 (%)	80.2	81.0	82.0	82.0	82.1	83.6	82.8	82.7	82.3	81.4

While evaluating the approach we observed that the main reason for errors of approach is related to segmentation of text lines and words to sub words. This problem is because of overlapping of bounding box of some originally disconnected sub words in vertical direction. Another problem is related to 'justifying' (Aligning text to both the left and right margins). This problem occurs in all typing softwares such as 'Microsoft Word'. In fact these softwares while justifying a Farsi or Arabic text, extend some words along their base lines; and therefore their width and consequently their aspect ratios become different with their original's. Another problem that can cause to error in proposed spotting method, is error in font recognition phase; because recognition rate of used method [11] for font recognition is 94.16% ; (5.84% error). In our method whatever the number of sub words of query word is more, performance of method (F1) will be better. In a comparison of other similar works which are about Arabic word spotting, our approach presents better result.

Table 2. Performance of two Arabic spotting methods and our method

Method	Precision	Recall	F1
Srihari [7]	55%	50%	0.524
Srihari [8]	60%	50%	0.545
Our method	92.3%	76.5%	0.836

After evaluating the font size recognition phase, we observed that our approach recognized font size of 471 document images out of 500 documents correctly and presented recognition rate of 94.2%.

4. Conclusion

In spite of important role of word spotting in information retrieval, we haven't found any paper about Farsi word spotting. Also there are only few papers about Farsi font recognition which all of them are font size independent; it means that they only recognize the font of document images but don't recognize their font sizes. This paper is the first paper about Farsi word spotting and font size recognition. In this approach when a Farsi keyword is entered by a user for searching in a document image, first, document's font face is recognized, then entered query word is rewritten and modified according to document's font face; and its image is constructed. In the next step, bounding box of modified word image is constructed and some font size independent features such as aspect ratio, mesh features, and number of its sub words are extracted for its description. These features are searched in the document image and the same word instances are found. After finding query word instances in document, with comparing width and height of entered keyword with width and height of found words from document, font size of document is determined. This approach has been evaluated on a dataset consisting of 500 Farsi document images which its font size recognition rate is 94.2% and its retrieval precision rate is 92.3% at recall rate of 76.5%.

References

1. Y. Pourasad, H. Hassibi, M. Banaeyan, International Review on Computers and Software (I.RE.CO.S). Vol. 6. No. 1 (2011) 55.
2. L. Shijian, T.L.Chew, Journal of Pattern recognition. No. 41 (2008) 1816.
3. L. Linlin, L. shijian, T.L. Chew, Ninth International conference on Document Analysis and Recognition (ICDAR). (2007).
4. L.Yue and C.L. Tan, International Journal of Pattern Recognition and Artificial Intelligence. Vol. 18, No. 2, (2004) 229.
5. H.K. Soo, C.P. Sang, B.J. Chang, S.K. Ji, H.R. Park, S.L. Guee, Digital Libraries, Implementing Strategies and sharing Experiences. Vol. 3815 (2005) 158.
6. S.N. Srihari, H. Srinivasan, P. Babu, C. Bhole, Proceedings of SPIE, San various scripts Jose, CA. (2006) 606702-1.
7. S.N. Srihari, H. Srinivasan, P. Babu, C. Bhole, Proceedings of Symposium on Document Image Understanding Technology (SDIUT-05), College Park, MD. (2005) 123.
8. S.N. Srihari, H. Srinivasan, C. Huang, S. Shetty, Indian Journal of Artificial Intelligence, Vol. 16, No. 3 (2006) 2.
9. R. Saabni, J. El-Sana, Proceedings of 11th International conference on frontiers on handwritten recognition (2008) 271.
10. R. Mehran, H. Pirsiavash, F. Razzazi, Techniques and Applications (DICTA 05). (2005) 385.
11. H. Khosravi, E. Kabir, Pattern Recognition Letters. 31 (2010) 75.
12. M. Zahedi, S. Eslami, Procedia Computer Science. 3 (2011) 1055.
13. M. Bahojb-Imani, M.R. Keyvanpour, R. Azmi, Procedia Computer Science. 3 (2011) 336.
14. S. Ben Moussa, A.Zahour, A. M. Benabdelhafid, A. Alimi, Pattern Recognition Letters. 5 (2010) 361.